A nonlinear mixed effects model to explain inter-individual variability in plant populations

Charlotte Baey[1], Samis Trevezas[1], and Paul-Henry Cournède[1]

Laboratory of Applied Mathematics and Systems, École Centrale Paris
Grande Voie des Vignes, 92290 Châtenay-Malabry
(E-mail: charlotte.baey@ecp.fr, samis.trevezas@ecp.fr, paul-henry.cournede@ecp.fr)

Abstract. It is common knowledge that the genetic variability of plants, even of the same variety, can be very important and, if we add locally varying climatic effects, the development of two neighboring similar plants could be highly different. This is one of the reasons why population-based methods for modeling plant growth are of great interest. A highly promising individual-based plant growth model is the GreenLab model [4] which was recently shown to have a good predictive capacity among competing models [1]. In this study, we extend the GreenLab formulation to the population level. In order to model the deviations from some fixed but unknown important biophysical and genetic parameters we introduce into the GreenLab model appropriate random effects. Under some assumptions, the resulting model can be cast into the framework of nonlinear mixed effects models. A stochastic variant of an EM-type algorithm (Expectation-Maximization) is generally needed to perform MLE for this type of incomplete data models and the interest is focused on the design of an efficient algorithm. In this direction, we propose a suitable Monte-Carlo EM (MCEM) algorithm for our model, where at each EM-iteration, MCMC is used to draw from the hidden states given the observed data. Data consist in organ mass measurements and are treated sequentially as first proposed in [11]. The performance of the algorithm is illustrated on simulated data from the sugar beet plant. Some possible extensions and improvements are also discussed.
Keywords: plant growth model, nonlinear mixed effects model, stochastic EM algorithm, MCMC methods, sugar-beet plant.

1 Introduction

Plants, as any other living organisms, are in constant interaction with each other. Genetic variability, even for plants of the same variety, as well as locally varying environmental conditions in a given field, can lead to the development of two different neighboring plants. This inter-individual variability can have a major impact at the agrosystem level, as shown for example by [3], who demonstrated that soil and crop micro-variability can have an impact on final yield, as some parts of the field can be more adapted to dryness, and can thus compensate less good performances of other parts of the field.

Individual-based plant growth models such as functional-structural plant models (FSPM) have gained a lot of success over the past years. These models describe the evolution of the 3D architecture of the plant over time, driven by the underlying ecophysiological processes (e.g., [13]), at the organ level. However, extrapolation to the field scale is still at its early stages. It mostly

concerns competition for light (e.g., [5]), and the calibration process is made from an average individual plant. If the level of description available in FSPMs made these approaches very appealing, their calibration on averaged individuals is not fully satisfactory as it only gives a partial representation of the field production.

In this study, we propose an extension of the individual-based Greenlab model, based on a bottom-up approach: the growth of each individual plant in a given field can be characterized by the same set of equations from the Greenlab model, but some of the model parameters are specific to this individual, and can therefore be considered as random effects. The resulting model can thus be cast into the framework of nonlinear mixed-effects models [7]. In this context, maximum likelihood estimators of the parameters can be obtained using an appropriate stochastic variant of an EM-algorithm (Expectation-Maximization) [10]. Due to the nonlinearity of the model, the E-step is in general analytically intractable, and an approximation of the Q-function should be done, but on the other hand, under suitable assumptions, the M-step can be resolved explicitly.

The methodology was developed here in the specific case of sugar beet crops, which have a very simple structure, as only three types of organs need to be considered (blades, petioles and root), but it can of course be applied to more complex plant structures, like maize or oilseed rape. The Greenlab model is introduced in Section 2.1, while the methodology is described in Section 2.3. Results from simulated data are presented in Section 3.

2 Material and Methods

2.1 The Greenlab model

The Greenlab model is a functional-structural model, combining rules for (i) biomass (mass for living organisms) production and allocation (functional part), and (ii) architectural development at the organ level (structural part). It was introduced by [8], and represented as a discrete dynamic system in [4]. Parameter estimation methods for this model are reviewed in [6]. Some recent advances for parameter estimation in the presence of modeling errors can be found in [11] and [12].

The first description of the Greenlab model as a discrete dynamic system was possible by taking advantage of the modular architectural development of plants. Indeed, the plant structure can be considered as the result of the accumulation of elementary botanical entities, called metamers, which usually correspond to some specific combinations of organs, characteristic for each plant species. The discretization of time is therefore possible by taking into account the time instants where metamers appear. The time interval between the appearance of two successive metamers is known as a growth cycle.

Despite the relative benefits of this inherent discretization, the choice of the growth cycle as a time step bears some limitations, especially for the functional part of the model and the handling of environmental data, since the latter are usually collected on a daily basis, while the growth cycle can vary from several days to one year in trees. To overcome these difficulties, a daily time step was

chosen to compute biomass production and allocation, but we still rely on the growth cycle for the creation of new organs.

In this paper we present in some detail the case of sugar beet which consists of three type of organs, that we denote by $\mathcal{O} = \{b, p, r\}$, where b stands for blade, p for petiole and r for root. Each blade and each petiole is defined by its rank, corresponding to the growth cycle at which it was preformed. The interest is focused on the functional part since the structural development is known and corresponds to the creation of one blade and one petiole at each growth cycle. The inter-individual variability of organogenesis has been studied by [2], who showed that it can be important in sugar beet populations.

Biomass production. The seed mass corresponds to the first biomass. After the appearance of the first leaf, biomass production is assumed to be given by :

$$F(t; p^*) = u_t \ \mu \ s^{pr} \left(1 - \exp\left(-k_b \ \frac{s^{act}(t; p_{al})}{s^{pr}}\right)\right), \tag{1}$$

where u_t stands for an environmental condition on day t (usually, the photosynthetically active radiation), s^{pr} an empirical coefficient related to the space occupied by the plant on the ground, μ an efficiency coefficient, $s^{act}(t; p_{al})$ the photosynthetically active foliar surface at the beginning of day t (see [11] for further details) depending on the allocation parameters p_{al} described in the next paragraph, and $p^* = (\mu, s^{pr}, k_b, p_{al})$.

Biomass allocation. A basic assumption of the Greenlab model is that biomass allocation to all expanding organs is proportional to organ specific functions, called sink functions and denoted by $s_{o,k}(u; p^o_{al})$. At a given time u, these functions depend on the type of the organ, and its expansion stage, i.e., the number of growth cycles that have elapsed since its creation. The basic factor determining the duration of a growth cycle, and consequently, organs demand for growth, is the temperature. For this reason it is very convenient to introduce the notion of thermal time, which is defined as follows :

$$\tau(u) = \int_0^u \max(0, T(s) - T_b) \ ds, \quad u \geq 0,$$

and represents at calendar time u, the accumulated sum of temperatures above a base temperature T_b until time u. In the sequel, for a leaf of rank k, we denote by τ_k its thermal time of initiation, τ^e_k its expansion period, and τ^s_k its lifetime. The thermal time of initiation of root is thus equal to τ_1, and we denote by τ^e_r its corresponding expansion period. We assume that root do not get senescent, and that initiation, expansion and lifetime of blades and petioles from the same leaf are identical.

After a first phase of initiation where the seed biomass is distributed uniformly in time, the produced biomass (due to photosynthesis) given by (1) is distributed to all expanding organs proportionally to

$$s_{o,k}(u; p^o_{al}) = c \ p_o \ \left(\frac{\tau(u) - \tau_k}{\tau^e_k}\right)^{a_o - 1} \left(1 - \frac{\tau(u) - \tau_k}{\tau^e_k}\right)^{b_o - 1} \mathbf{1}_{\tau_k \leq \tau(u) \leq \tau_k + \tau^e_k},$$

where $p^o_{al} = (p_o, a_o, b_o)$ for $o \in \mathcal{O}$ and c is the normalizing constant of a discrete beta law B(a_o,b_o).

The sum of all sink functions on day u defines the total biomass demand $d(u; p^{al})$ on day u, and the ratio $s_{o,k}(u; p^o_{al})/d(u; p^{al})$ determines the percentage of the produced biomass $F(t; p^*)$ which is allocated to the organ of type o and rank k at the end of day u.

2.2 A two-stage formulation of the model

To account for inter-individual variability, random effects are introduced in the Greenlab model, which can then be seen as a two-stage hierarchical one.

First-stage: intra-individual variation. We denote by $\bar{z} = (\bar{z}_{i,n})_{1\leq i\leq s,\ 0\leq n\leq n_i}$ the theoretical biomasses of organs of rank $n+1$ for plant i. The theoretical biomasses $\bar{z}_{i,n}$ can be obtained as a function of the sequence of produced biomasses:

$$\bar{z}_{i,n} = G_n(\phi_i) = \left(\sum_{\tau(t)=\tau_{n+1}}^{\tau^e_k \wedge \tau_{max}} \frac{s_{o,n}(t; p^{al})}{d(t; p^{al})}\ F(t; \phi_i) \right)_{o\ \in \mathcal{O}}, \tag{2}$$

where ϕ_i is the vector of parameters specific to plant i, G_n is the vector-valued function of the theoretical biomasses of organs of rank $n+1$, and τ_{max} is the thermal time at which observations are made.

To account for positivity in mass measurements we define $\bar{y}_{i,n} = \log(\bar{z}_{i,n})$. If we denote by $y = (y_{i,n})_{1\leq i\leq s,\ 0\leq n\leq n_i}$ the vector of mass measurements in the log-scale and by $\Sigma_{b,p}$ and σ^2_r a two-dimensional covariance matrix and variance parameter respectively, then we assume that :

$$y_{i,n} = \bar{y}_{i,n} + \epsilon_{i,n}, \quad \epsilon_{i,n} \sim \mathcal{N}_{d_n}(0, \Sigma_n),\ 1 \leq i \leq s,\ 0 \leq n \leq n_i, \tag{3}$$

where $(\epsilon_{i,n})_{1\leq i\leq s,\ 0\leq n\leq n_i}$ are mutually independent random variables and

$$\Sigma_n = \begin{cases} \text{diag}\{\Sigma_{b,p}, \sigma^2_r\} & \text{if } n = 0, \\ \Sigma_{b,p} & \text{if } n \geq 1. \end{cases} \tag{4}$$

In this way, the measurement errors from organs of two different plants or of the same plant but with different ranks are assumed to be independent.

Second-stage: inter-individual variation. In this second stage, the variability of the subject-specific parameters defined in the previous stage, ϕ_i, is assessed. We assume the following model for the vector $\phi_i = (\phi_{i,1}, \ldots, \phi_{i,P})^t$, with P the number of random parameters:

$$\begin{aligned} &\phi_i = \beta + \xi_i, \\ &\xi_i \sim \mathcal{N}_P(0, \Gamma), \end{aligned} \tag{5}$$

where β is the vector of fixed effects and Γ a diagonal covariance matrix.

2.3 Parameter estimation

We denote by $\theta = (\theta_1, \theta_2)$, where $\theta_1 = (\beta, \sigma_1^2, \ldots, \sigma_P^2)$ and $\theta_2 = \Sigma_{b,p}$, the vector of unknown parameters. To compute the maximum likelihood estimator of θ, we need to compute the likelihood of the model, which will be in general analytically intractable due to the nonlinearity of G_n given by (2). However, our model can be seen as an incomplete data model, with $y = (y_{i,n}, 1 \leq i \leq s,\ 0 \leq n \leq n_i)$ the observed data, the random effects $\phi = (\phi_i, 1 \leq i \leq s)$ being the unobserved data. The complete data of the model is (y, ϕ), and in such cases, an appropriate variant of an EM-algorithm [9] (Expectation-Maximization) can be implemented to approximate the maximum likelihood estimator of θ.

Each iteration of this algorithm consists in two steps: the expectation step (E-step) in which the conditional expectation of the complete data log-likelihood given the observed data is computed under the current parameter value, and the maximization step (M-step) in which the parameters are updated by maximizing the Q-function obtained in the E-step. The two steps of the EM-algorithm are described below:

E-step. At iteration k, the E-step of the algorithm consists in the computation of the Q-function given the current value of the parameter θ^k. Due to the independence between plants and between organs of different ranks within the same plant, the Q-function can be decomposed as follows:

$$\begin{aligned} Q(\theta; \theta^k) &= \sum_{i=1}^{s} \mathbb{E}_{\theta^k}(\log f(\phi_i; \theta_1) \mid y) + \sum_{i=1}^{s} \sum_{n=0}^{n_i} \mathbb{E}_{\theta^k}(\log f(y_{i,n} \mid \phi_i; \theta_2) \mid y) \\ &= Q_1(\theta_1; \theta^k) + Q_2(\theta_2; \theta^k). \end{aligned} \tag{6}$$

M-step. In the M-step of the algorithm, we maximize the Q-function with respect to θ. Thanks to the decomposition of the Q-function given by (6), maximizing $Q(\theta; \theta^k)$ with respect to θ is equivalent to maximizing $Q_1(\theta_1; \theta^k)$ with respect to θ_1 and $Q_2(\theta_2; \theta^k)$ with respect to θ_2. The update equations can be obtained easily and are given by:

$$\hat{\beta}_j = \frac{1}{s} \sum_{i=1}^{s} \mathbb{E}_{\theta^k}(\phi_{i,j} \mid y_i), \tag{7}$$

$$\hat{\sigma}_j^2 = \frac{1}{s} \sum_{i=1}^{s} \mathbb{E}_{\theta^k}(\phi_{i,j}^2 \mid y_i) - \hat{\beta}_j^2, \tag{8}$$

$$\hat{\Sigma}_{b,p} = \frac{1}{s + \sum_{i=1}^{s} n_i} \sum_{i=1}^{s} \sum_{n=0}^{n_i} \mathbb{E}_{\theta^k} \left[(y_{i,n} - \log G_n(\phi_i)) (y_{i,n} - \log G_n(\phi_i))^t \mid y_i \right], \tag{9}$$

where $y_{i,0}$ is restricted to blade and petiole only.

From these equations, we can see that at each iteration of the EM-algorithm, the problem of maximization is reduced to the problem of computing conditional expectations given y_i, under the current parameter value θ^k.

Approximation of the E-step Due to the nonlinearity of the model the E-step given by (6) cannot be performed explicitly. However, many stochastic variants of the EM algorithm are available to approximate a non-explicit E-step. In this paper, as a first implementation we tried a Monte Carlo EM (MCEM, [14]) algorithm, where the Q-function is approximated via Monte Carlo simulations. In particular, hidden data are drawn via an MCMC (Markov Chain Monte Carlo) algorithm like Metropolis-Hastings or Gibbs Sampling. At iteration k of the algorithm, we simulated for each individual plant a Markov Chain of size M, with stationary distribution our target distribution $f(\phi_i \mid y_i; \theta^k)$.

3 Results

The methodology was first applied to a set of 50 simulated plants. Thanks to a preliminary sensitivity analysis, the two most influential parameters were shown to be μ and s^{pr}. Consequently, as a first approach, random effects were only used for these parameters. The other parameters were assumed to be known. Concerning s^{pr}, we assumed that $\log s^{pr} \sim \mathcal{N}(\beta_2, \sigma_2^2)$. Finally, $\theta_1 = (\beta_1, \beta_2, \sigma_1, \sigma_2)$ and $\theta_2 = (\sigma_b^2, \sigma_p^2, \rho)$, where the latter vector corresponds to the variance parameters and the correlation coefficient of the covariance matrix $\Sigma_{b,p}$ (see, (4)). In the tests that we present, the parameter σ_r^2 was fixed.

In Table 1 we present the parameter estimation results that we obtained. We used two different initializations and three independent runs for each one of them. At each iteration, a Markov chain of size 1000 was generated for each plant. The proposal distribution was set equal to the prior distribution of the hidden data under the current parameter value. For these tests, the algorithm stopped manually after 60 iterations. Table 1 gives as well the values that we used to generate the data, and the MLE if the data were fully observed. Two different sets of initial values were tested: for the first three runs (columns 4 to 6), the initial values were 4.5 for μ and -4 for s^{pr}, and for the last three runs (columns 7 to 9) the initial values were 5.5 for μ and -2 for s^{pr}.

Parameter	True value	Fully-observed	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6
β_1	5.5	5.514	5.512	5.508	5.507	5.510	5.509	5.510
β_2	-3	-2.979	-2.980	-2.980	-2.980	-2.980	-2.980	-2.980
σ_1	0.1	0.101	0.068	0.070	0.074	0.067	0.070	0.066
σ_2	0.1	0.089	0.083	0.084	0.085	0.084	0.083	0.085
σ_b^2	0.1	0.100	0.1135	0.114	0.114	0.114	0.113	0.113
σ_p^2	0.1	0.100	0.115	0.115	0.115	0.115	0.115	0.115
ρ	0	-0.004	0.009	0.010	0.010	0.010	0.009	0.009

Table 1. Parameter estimation results.

Figure 1 shows that the convergence was reached quickly for β_1 and β_2, but more iterations are needed for the variance of the random effects, especially for σ_1. The results from different initializations and independent runs were also very encouraging (Table 1), even for the estimations of observation noises.

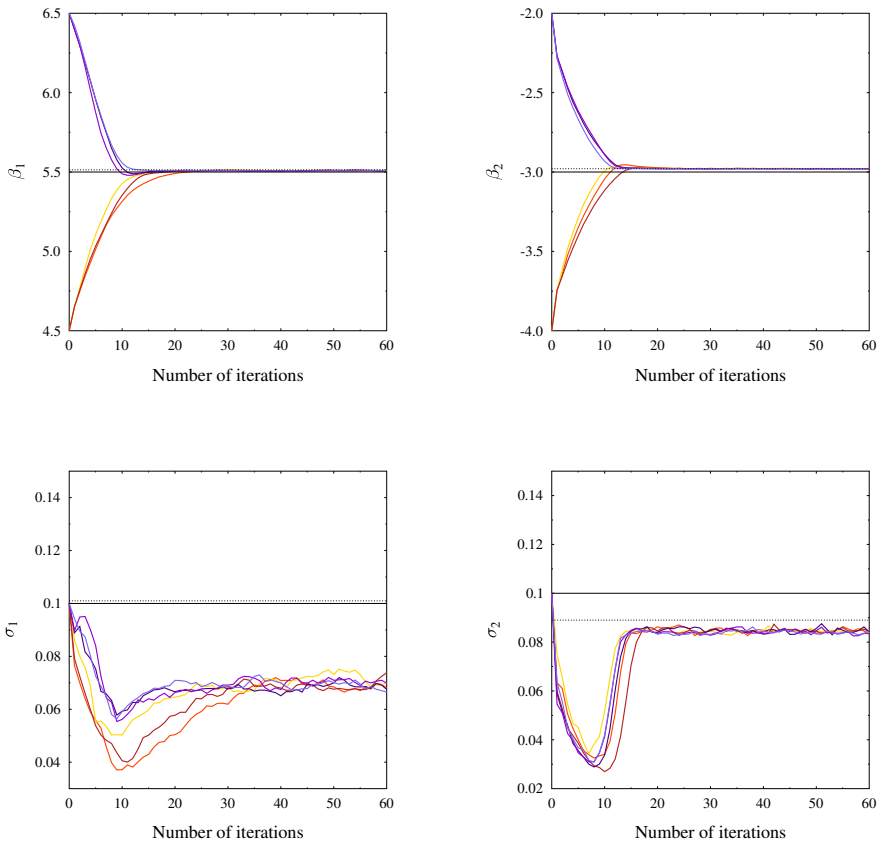

Fig. 1. Parameter estimation for 60 iterations of the Metropolis-Hastings algorithm: (top) estimation of β_1 and β_2, (bottom) estimation of σ_1 and σ_2. The solid lines represent the values used to generate the data and the dotted line correspond to the maximum likelihood estimators if the data were fully observed.

4 Discussion

The methodology that we presented in this paper to account for inter-individual variability in plant populations is suitable for a large number of crop plants. Results from simulated data were encouraging, and we are currently working with real data and with more random parameters. Different MCMC versions of the current MCEM algorithm will be compared in the sequel and the latter will be compared with a Stochastic Approximation EM (SAEM, see [10]) algorithm. An automated MCEM algorithm can easily be implemented (see, [12]). The latter paper presents a parameter estimation method by including modeling errors in biomass production for a single plant. As a future step, we will extend our population based approach to account for modeling errors as well.

Finally, it is noteworthy that with the proposed methodology, approximated confidence intervals can be easily obtained as a by-product of the algorithm.

References

1.C. Baey, A. Didier, L. Song, S. Lemaire, F. Maupas and P.-H. Cournède, Evaluation of the Predictive Capacity of Five Plant Growth Models for Sugar Beet. *Proc. of 4th Int. Symp. on Plant Growth Modeling, Simulation, Visualization and Application.* Shanghai, 2012.

2.C. Baey, A. Didier, S. Lemaire, F. Maupas, P.-H. Cournède. Modelling the inter-individual variability of organogenesis in sugar beet populations using a hierarchical segmented model. *Ecological Modelling* (to appear), 2013.

3.J. Brouwer, L. K. Fussell, and L. Herrmann. Soil and crop growth micro-variability in the West African semi-arid tropics: a possible risk-reducing factor for subsistence farmers. *Agriculture, Ecosystems and Environment*, 45, 3-4, 229–238, 1993.

4.P.-H. Cournède, M.Z. Kang, A. Mathieu, J.-F. Barczi, H.P. Yan, B.G. Hu, and P. de Reffye. Structural factorization of plants to compute their functional and architectural growth. *Simulation*, 7, 82, 427–438, 2006.

5.P.-H. Cournède, A. Mathieu, F. Houllier, D. Barthélémy, P. de Reffye. Computing competition for light in the GreenLab model of plant growth: a contribution to the study of the effects of density on resource acquisition and architectural development. *Annals of Botany*, 8, 101, 1207–1219, 2008.

6.P.-H. Cournède, V. Letort, A. Mathieu, M.-Z. Kang, S. Lemaire, S. Trevezas, F. Houllier, and P. de Reffye. Some parameter estimation issues in functional-structural plant modelling, *Mathematical Modelling of Natural Phenomena*, 6, 133-159, 2011.

7.M. Davidian and D. Giltinan. *Nonlinear Models for Repeated Measurement Data*, Chapman and Hall, New-York, 1995.

8.P. de Reffye and B.-G. Hu. Relevant qualitative and quantitative choices for building an efficient dynamic plant growth model: Greenlab case. *Proc. of the 2nd Int. Symp. on Plant Growth Modeling, Simulation, Visualization and Their Applications*, Beijing, China, 2003, IEEE, 87–197

9.A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B.*, 39, 1, 1–38, 1977.

10.E. Kuhn, and M. Lavielle. Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics and Data Analysis*, 49, 1020-1038, 2005.

11.S. Trevezas, and P.-H. Cournède. A sequential Monte Carlo approach for MLE in a plant growth model. *Journal of Agricultural, Biological, and Environmental Statistics*, (to appear), 2013.

12.S. Trevezas, S. Malefaki and P.-H. Cournède. Simulation techniques for parameter estimation via a stochastic ECM algorithm with applications to plant growth modeling, preprint, 2013.

13.J. Vos, L.F.M. Marcelis, P. de Vissers, P. Struik and J.B. Evers. *Functional-Structural plant modeling in crop production.* Springer, Berlin, 2007.

14.G. C. G. Wei, and M. A. Tanner. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85, 411, 699–704, 1990.